JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 10: 1441~1446, Oct. 2022

사용자 관점에서의 음식 레시피 분류 모델에 관한 연구

이우행 · 최수연*

Food Recipe Clustering Model from the User's Perspective

Woo-Hang Lee[1] · Soo-Yeun Choi[1*]

*Graduate Student, Computer & Information Technology, Korea University Graduate School, Seoul, 02841 Korea

요 약

현대인들은 음식 레시피에 대한 다양한 정보들을 인터넷이나 소셜 미디어 등에서 매우 쉽게 접할 수 있게 되었다. 음식 레시피를 제공하는 공급량이 많아지면서 범람하는 정보 안에서 사용자들이 각자에 맞는 레시피를 찾기에는 수고로움이 따르게 된다. 이처럼 사용자들의 요구사항을 반영하여 정보를 제공할 필요성이 높아졌고, 음식 레시피와 요리 추천에 관련 연구가 활발해지고 있다. 또한, 이를 활용한 인터넷, 영상 및 어플리케이션 시장 역시 급속도로 활성화되고 있다. 본 연구에서는 음식 레시피 사용자들의 관점에서 레시피를 분류하기 위하여 사용자의 리뷰 데이터를 비지도학습인 K-평균 군집화 기법을 적용하였으며, 이를 통해 "음식 레시피 분류 모델"을 도출하였다. 그 결과 특정 목적, 조리 단계 등 많은 사용자들이 필요한 정보를 포함한 총 25개의 군집으로 분류하였다.

ABSTRACT

Modern people can access various information about food recipes very easily on the Internet or social media. As the supply of food recipes increases, it is difficult to find a suitable recipe for each user in the overflowing information. As such, the need to provide information by reflecting users' requirements has increased, and research related to food recipes and cooking recommendations is becoming active. In addition, the Internet, video, and application markets using this are also rapidly activating. In this study, in order to classify recipes from the user's perspective of food recipe users, the user's review data was applied with the k-mean clustering technique, which is unsupervised learning, and a "food recipe classification model" was derived. As a result, it was classified into a total of 25 clusters including information needed by many users, such as specific purposes and cooking stages.

키워드 : k-평균 군집화, 분류, 비지도학습, 음식, 레시피

Keywords : k-means Clustering, Classification, Unsupervised Learning, Food, Recipe

Received 12 September 2022, **Revised** 15 September 2022, **Accepted** 17 September 2022

*** Corresponding Author** Soo-Yeun Choi(E-mail:sooyeun1202@naver.com, Tel***－****－****)
Graduate Student, Computer & Information Technology, Korea University Graduate School, Seoul, 02841 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.10.1441 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 1인 가구가 증가하면서 혼밥 문화, 먹방과 쿡방 등 요리 관련 프로그램과 개인 방송들이 꾸준히 인기를 얻고 있다. 과거와 비교했을 때, 현대인들은 음식에 대한 다양한 정보들을 인터넷, 소셜 미디어 등을 통해 보다 쉽게 접할 수 있게 되었다. 또한 블로그, 트위터, 페이스북, 인스타그램, 유튜브 등 다양한 플랫폼에서는 요리 관련 데이터와 레시피 정보가 범람하고 있다. 하지만, 그 중에서 사용자의 목적이나 취향에 적합하거나 유의미한 정보들을 선별하여 얻기는 쉽지 않다. 따라서 본 연구에서는 정보 제공자가 아닌 사용자가 원하는 레시피 분류 체계를 제공하기 위하여 레시피 태그 데이터 분석을 실시하여, 사용자의 관점에서의 레시피 분류 방법을 제안하고자 한다.

Ⅱ. 본 론

2.1. 관련 연구

IT기술을 활용한 레시피 관련 연구는 크게 두 가지의 관점으로 나뉜다. 첫 번째는 레시피의 텍스트 분석에 초점을 맞춘 연구이다. 해당 연구에서는 레시피 분류 모델은 재료와 조리 동작을 중심으로 텍스트 분석 기법[1-3]을 사용하거나 요리 난이도 및 재료의 유사성, 맛을 기준으로 텍스트를 분석하는 기법[4-5]을 사용한다. 두 번째로는 사용자의 정보들을 수집하여 레시피를 추천하는 것에 초점을 맞춘 연구로, 협업 필터링을 이용하여 레시피를 추천하는 시스템[6]에 대한 연구가 진행되었다.

이와 같은 선행 연구들은 재료 중심으로 텍스트 분석에 초점을 맞출 경우 사용자의 요리 실력이나 원하는 취향을 고려하지 못하는 한계점이 있으며, 레시피 추천에 초점을 맞출 경우 사용자 개인의 데이터를 수집하지 못할 경우 연구를 진행하기 어려움이 있다.

본 연구는 요리 레시피에 포함된 데이터 분석을 통해 사용자 관점에서의 유용한 분류를 도출할 수 있는 방법을 제안하였다. 본 연구에서는 레시피 사용자가 작성한 태그 등의 텍스트 데이터를 정형화 하고 데이터 간 의미적 유사성을 계산하여 활용하였다.

2.2. 이론적 배경

2.2.1. TF-IDF(Term Frequency - Inverse Document Frequency) [7]

TF-IDF(Term Frequency-Inverse Document Frequency)는 정보 검색과 텍스트마이닝에서 이용하는 가중치로, 여러 문서로 이루어진 문서 군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. TF(단어 빈도, term frequency)는 특정한 단어가 문서 내에 얼마나 등장하는지를 나타내는 값, DF(문서 빈도, document frequency)는 단어 자체가 문서군 내에서 등장하는 빈도이며, 이 값의 역수를 IDF(역문서 빈도, inverse document frequency)라고 한다. TF-IDF는 TF와 IDF를 곱한 결과 값을 의미한다. TF-IDF는 아래와 같은 식으로 나타낸다.

$$TF = \frac{n}{N} \tag{1}$$

$$IDF = -\log_{10}\frac{d}{D} = \log_{10}\frac{D}{d} \tag{2}$$

$$TF-IDF = TF \cdot IDF = \frac{n}{N} log_{10}\frac{D}{d} \tag{3}$$

N: 단어의 개수, D: 문장의 개수
n:특정 단어의 개수, d: 특정 문장의 개수

2.2.2. K-means[8]

K-means 클러스터링은 비지도 학습(Unsupervised Learning)의 한 방법으로, 처음에 랜덤으로 k 개의 클러스터의 중심점을 선택하고 각각의 점들을 가장 가까운 중심점으로 할당한 후 평균을 구해 중심점을 바꾸는 과정을 클러스터의 중심점이 변하지 않을 때까지 반복하여 군집화를 시키는 방법을 의미한다. K-means는 비지도 학습이기 때문에 정답의 데이터 셋을 요구하지 않고 별도의 정보 없이 데이터를 분류할 수 있으며, 대량의 데이터를 빠르게 처리하는데 용이하다.

2.3. 데이터 수집 및 분석

2.3.1. 데이터 수집

일반적으로 레시피란 식재료를 음식으로 만드는 방법이며, 광의적으로 재료의 종류 및 선택, 조리에 필요

한 도구, 구체적인 조리법, 데코레이션 등 요리에 필요한 모든 사항을 포함한다. 본 연구에서 '레시피'는 광의적 개념으로써, 요리에 사용한 재료, 소요 시간, 요리 단계 등을 모두 포함한다.

사용자 관점에서의 레시피 정보를 얻기 위해 Kaggle에서 18년간 Food.com(formely GeniusKitchen)에서 업로드한 데이터 셋을 사용했다. 해당 데이터에는 레시피와 사용자의 태그 데이터를 포함하고 있다. 레시피 데이터는 음식명(name), 음식 고유번호(id), 소요 시간(minutes), 사용자의 태그(tags), 재료명(ingredients) 등 12개의 속성이 있으며, 사용자의 태그 데이터에는 사용자 번호, 음식 고유번호, 평점, 태그 등의 정보를 포함하고 있다. 활용한 데이터 셋(Recipe Data)에 포함된 12개의 속성은 아래 표. 1과 같다.

Table. 1 Recipe Data

raw-data list	description
name	recipe name
id	recipe unique id
minutes	cooking time
contributor_id	contributor unique id
submitted	submitted date
tags	recipe tag
nutrition	nutrition name
n_steps	Number of cooking steps
steps	cooking steps
description	recipe detail desctiption
ingredients	ingredients
n_ingredients	Number of ingredients

각각 5,000개의 샘플 데이터를 추출하여 음식명, 조리 단계, 음식의 상세 설명, 재료명, 사용자의 태그 데이터 5가지 속성으로 분류를 진행하였다. 5가지 속성에 따른 분류 중, 요리명은 기존 분류 체계와 동일한 형태를 띠었고, 그 외에 속성 중에서 가장 유의미한 분류체계인 태그 데이터 분석을 통해 사용자 관점에서의 분류를 진행한다.

2.3.2. 데이터 분석

사용자의 태그 데이터가 포함된 레시피 수는 총 231,637개로 구성되어 있으며, 결측 값의 개수는 4,980개이다. 결측 값을 제거한 뒤 기초 통계 분석을 실시하였으며, 전체 사용자의 태그 데이터의 상위 10개의 빈도수는 그림. 1과 같다. 사용자의 태그 데이터에 많이 내포되어 있는 정보는 요리의 시간(time-to-make, 60-minutes-or-less), 요리의 분류(course, main-dish), 요리의 난이도(easy) 등이다. 또한, 하단의 그림. 2를 살펴보면 easy, pork, peanut butter, apple, pot 등의 재료명, 난이도, 조리 도구 등의 단어들이 나타나는 것을 알 수 있다.

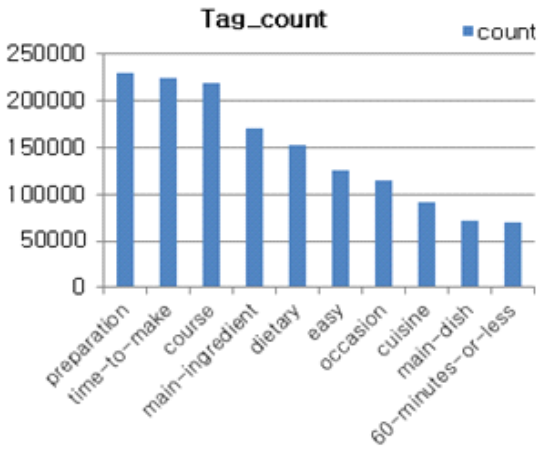

Fig. 1 Tag Content Frequencies(Top 10)

Fig. 2 Word Cloud of the Entire Tag Data

2.3.3. 실험 모델

본 연구에서는 효율적으로 대용량의 자연어를 분석하기 위하여 K-means 기법을 활용하여 레시피에 대한 클러스터링을 진행하였다. 사용자의 태그 데이터의 클러스터 개수를 정하기 위해 그래프를 그리면 그림. 3과 같다. 해당 그래프를 살펴보면 기울기의 변화가 달라지는 구간이 클러스터 개수가 25, 50개 근처에 존재한다. 클러스터의 수가 증가하면 분류는 정확해지지만 의미있는 분류를 하기 어려워질 수 있다. 따라서 해당 연구에서는 클러스터의 개수를 25개로 선정하여 클러스터링을 진행한다.

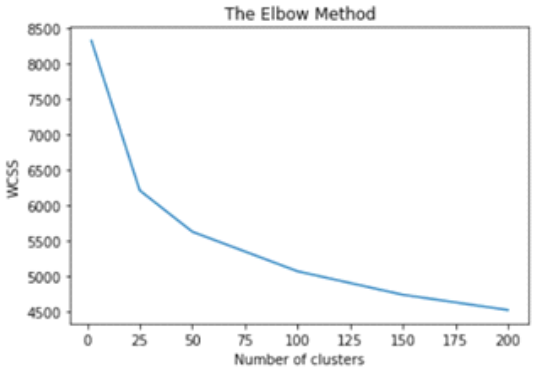

Fig. 3 Determine the number of clusters

사용자의 태그 데이터를 25개의 클러스터로 분류했을 때 분류의 대표명은 표. 2와 같다. 해당 분류는 soup, pasta, pie, salad와 같은 요리 종류, appetizer, side, dish 등의 요리 코스 단계, pork, bread, beef, egg 등의 재료 등의 다양한 분류를 내포한다. 이는 대표적으로 사용자들이 많이 이용하는 정보를 의미하며, 이를 통하여 새로운 시각의 재분류 관점도 생성할 수 있다.

Table. 2 25 Clusters based on Tag Data

Cluster number	Category name
Cluster 0	soup
Cluster 1	pasta
Cluster 2	occasion
Cluster 3	large
Cluster 4	fat
Cluster 5	appetizer
Cluster 6	pork
Cluster 7	bread
Cluster 8	free
Cluster 9	sauce
Cluster 10	make
Cluster 11	pie
Cluster 12	american
Cluster 13	high
Cluster 14	side
Cluster 15	dish
Cluster 16	salad
Cluster 17	beef
Cluster 18	egg
Cluster 19	healthy
Cluster 20	chicken
Cluster 21	fish
Cluster 22	cooky
Cluster 23	beverage
Cluster 24	dessert

위에서 분류한 25개 중 대표로 클러스터 0과 클러스터 22를 살펴보기 위하여, 각 클러스터 안에 있는 태그 데이터를 워드 클라우드로 나타내었다. 해당 워드 클라우드는 각각 그림. 4, 그림. 5와 같이 나타난다. 클러스터 r0인 soup에는 potato, bean, stew, creamy 등의 단어가 내포된 것을 알 수 있으며, 수프와 관련된 재료와 식감을 많이 언급하는 것을 알 수 있다. 또한, 클러스터 22인 cooky에는 chocolate chip, brownie, peanut butter, square 등의 단어들이 내포되어 있으며, 쿠키와 관련된 재료와 형태가 많이 언급되는 것을 알 수 있다. 이와 같이 사용자들이 필요한 정보를 포함하고 있기 때문에 사용자 관점에서의 레시피 분류명과 분류 내 포함된 정보들을 확인할 수 있다.

Fig. 4 Word Cloud of Cluster 0

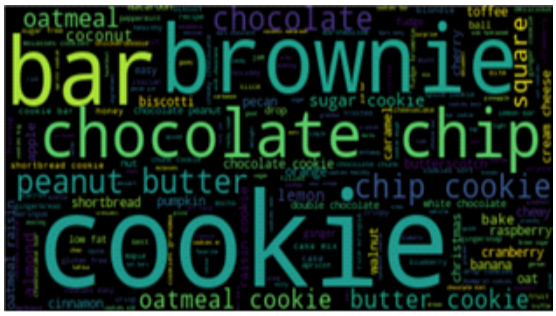

Fig. 5 Word Cloud of Cluster 22

Ⅲ. 결 론

본 연구에서는 비지도 학습인 K-means 활용하여 사용자 태그 데이터 내 텍스트 유사성을 통해 사용자 관점에서의 레시피 분류를 제안하였다. 사용자의 태그 데이터에는 재료 외에 소요시간과 국가, 특정 목적, 조리 단계 등 많은 사용자들이 필요한 정보를 포함하고 있기 때문에 기존에 제안되었던 재료에 대한 분류와의 차별성이 있다. 또한, 25개 분류 안에서도 재료 명, 음식 코스 단계, 건강, 요리 명 등의 분류 등 대분류 및 새로운 기준의 분류 체계도 제안할 수 있다.

한편 본 연구는 몇 가지 한계점을 가지고 있다. 첫째, 특정 사이트의 사용자 태그 정보를 사용하였기 때문에 데이터의 표본이 편향되어 있다. 둘째, 사용자의 태그 데이터는 의도적으로 많이 사용되는 단어를 임의적으로 조작하여 작성할 수 있다. 따라서 태그라는 하나의 요소만으로 분류체계를 정의하는 것을 편향되고 조작된 결과를 가지고 올 수 있다. 따라서, 향후에는 태그 이외에 정보들을 함께 고려할 수 있는 부분에 대한 연구가 진행되어야 한다. 본 연구는 사용자들의 관점에서 레시피 분류를 수행하고 업데이트된 데이터를 반영하여 새로운 음식 레시피 분류를 제시하여 실제 사용자들이 유익한 정보를 얻을 것이라 기대한다.

음식, 요리에 대한 관심이 점차 높아짐에 따라 소셜미디어, 인터넷 등 여러 매체를 통한 레시피 관련 데이터는 지속적으로 증가할 것이다. 이에 레시피 제공자들은 사용자를 고려한 체계적인 분류를 만들고 효율적으로 시스템을 운영할 필요성이 증가할 것이다. 본 연구기반으로 사용자 관점에서의 음식 레시피를 분류하고 더 나아가 개인별 맞춤 분류 제공 등에도 활용할 수 있을 것으로 기대한다.

References

[1] J. Hong and H. Lee, “Recipe Recommendation Method Using Text Analytics and Ingredients Hierarchy,” *Journal of the Korean Institute of Industrial Engineers*, vol. 45, no. 4, pp. 302-312, Aug. 2019.

[2] J. Choi and G. Han, "Structural Analysis of Cooking Recipe Texts - Based on Kimchi Jjigae Recipe," *The Korean Journal of Community Living Science*, vol. 28, no. 2, pp. 191-201, May 2017.

[3] D. Lee, I. Min, J. -W. Kim, J. Lee, J. Shin, and S. Lee, "Design and Implementation of Food Recommendation System Based on Personal Preference," in *Proceedings of the 2016 Winter Conference of the Korean Institute of Information Scientists and Engineers*, Pyeongchang, Korea, pp. 1411-1413, 2016.

[4] J. Mueller and A. Thyagarajan, "Siamese Recurrent Architectures for Learning Sentence Similarity," in *Proceedings of the National Conference on Artificial Intelligence*, Phoenix: AZ, USA, vol. 30, no. 1, pp. 2786-2792, 2016.

[5] D. H. Kim, “Comparison of Taste Prediction Performance of Recipe using Machine learning,” in *Proceedings of Symposium of the Korean Institute of communications and Information Sciences*, Jeongseon, Korea, pp.1088-1090, 2018.

[6] J. H. Jo, "Development of wine recommendation algorithm using similarity algorithm -Focus on Bigdata analysis techniques," M.S. thesis, Namseoul University, 2018.

[7] M. Dillon, “Introduction to modern information retrieval,” *Information Processing & Management*, vol. 19 no. 6, pp. 402-403, 1983.

[8] S. -J. Choi, “Beta-wave Correlation Analysis Model based on Unsupervised Machine Learning,” *Journal of Digital Convergence*, vol. 17, no. 3, pp. 221-226, Mar. 2019.

이우행(Woo-Hang Lee)

빅데이터융합과 석사
※관심분야 : 텍스트 마이닝, 빅데이터 분석 및 예측 모델링, 데이터 시각화

최수연(Soo-Yeun Choi)

빅데이터융합과 석사
※관심분야 : 텍스트 마이닝, 빅데이터 분석 및 예측 모델링, 데이터 시각화